J. Phycol. **42,** 78–85 (2005)

DOI: 10.1111/j.1529-8817.2005.00164.x

ANALYSIS OF EXPRESSED SEQUENCE TAGS (ESTS) FROM THE POLAR DIATOM *FRAGILARIOPSIS CYLINDRUS*[1]

Thomas Mock[2]

Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany

Andreas Krell

Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany

Gernot Glöckner

Genome Analysis, Institute for Molecular Biotechnology, Beutenberg Str. 11, 07745 Jena, Germany

Üner Kolukisaoglu

University of Rostock, Department of Plant Physiology, Albert-Einstein-Str. 3, 18051 Rostock, Germany

and

Klaus Valentin

Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany

Analysis of expressed sequence tags (ESTs) was performed to gain insights into cold adaptation in the polar diatom *Fragilariopsis cylindrus* Grunow. The EST library was generated from RNA isolated 5 days after *F. cylindrus* cells were shifted from approximately $+5^{\circ}$ C to -1.8° C. A total of 1376 ESTs were sequenced from a non-normalized cDNA library and assembled into 996 tentative unique sequences. About 27% of the ESTs displayed similarity (tBLASTX, *e*-value of $\leq 10^{-4}$) to predicted proteins in the centric diatom *Thalassiosira pseudonana* Hasle & Heindal. Eleven additional algae and plant data bases were used for annotation of sequences not covered by *Thalassiosira* sequences (7%). Most of the ESTs were similar to genes encoding proteins responsible for translation, ribosomal structure, and biogenesis (3%), followed by genes encoding proteins for amino acid transport and metabolism and post-translational modifications. Interestingly, 66% of all the EST sequences from *F. cylindrus* displayed no similarity (*e*-value $\leq 10^{-4}$) to sequences from the 12 non-redundant databases. Even 6 of the 10 strong to moderately expressed sequences in this EST library could not be identified. Adaptation of *F. cylindrus* to freezing temperatures of seawater may require a complex protein metabolism and possibly also genes, which were highly expressed but still unknown. However, it could also mean that due to low temperatures, there might have been a stronger pressure to adapt amino acid sequences, making it more difficult to identify these unknown sequences and/or that there are still few protist sequences available for comparison.

Key index words: **EST; cold acclimation; diatom; gene expression; polar; genome**

Abbreviations: **cDNA, complementary DNA; *e*, expectancy value; EST, expressed sequence tag; fcp, fucoxanthin/chl binding protein; TC, tentative consensus**

The polar seas are one of the least studied and least understood ecosystems on the planet. Photosynthesis and oxygen production in polar regions are mainly accomplished by single-celled phytoplankton that live in the top portion of the ocean and within sea ice (Legendre et al. 1992, Lizotte 2001, Boyd 2002, Thomas and Dieckmann 2002). Relatively few classes of phytoplankton and a small number of genera and species appear to play key roles in trophic interactions and biogeochemical fluxes such as that of carbon (Falkowski et al. 2004). The most important class of phytoplankton in the polar marine food-web are diatoms (Bacillariophyceae, Smetacek 1999), with many species that are highly stenothermal with upper lethal temperature limits of only approximately $+10^{\circ}$ C (Fiala and Oriol 1990). Polar diatoms are well adapted to the main stressors in polar seas: constant low and freezing temperatures, physical disturbances from sea ice, and extreme seasonality (Cota 1985). However, they probably benefit from the relatively high concentrations of dissolved silicate ($Si(OH)_4$) in polar seawater. Species prominent in sediments also serve as proxies in the reconstruction of paleoclimate (Leventer 1998). In the past, because of the

[1]Received 7 September 2004. Accepted 31 August 2005.
[2]Author for correspondence and present address: School of Oceanography, University of Washington, Box 357940, Seattle, WA 98195, USA. email mockt@u.washington.edu.

important global role of diatoms in polar marine ecosysterms, many investigations were carried out using *in situ* perturbation experiments (e.g. EISENEX, EIFEX, SOFeX, ISPOL), meso- and microcosm studies, and physiological and biochemical experiments to assess their adaptation and to understand their role in the biogeochemical cycles of polar oceans (Boyd 2002, Coale et al. 2004, Falkowski and Davis 2004). Despite these efforts, it is still unclear what features allow polar diatoms to survive under polar conditions.

Genome sequencing has become a powerful tool and many genomes from key organisms have been sequenced to provide the basis for understanding biochemical and physiological activities in these organisms. Early reports of genome sequences from autotrophic prokaryotes by Rocap et al. 2003, Palenik et al. 2003, along with a paper by Dufresne et al. 2003, demonstrate how genomic studies can lead to a new understanding of biodiversity, ecology, biological efficiency, and biogeochemistry in marine systems. Recently, the first complete genome of an ecologically important marine diatom (*Thalassiosira pseudonana*) was sequenced (Armbrust et al. 2004). New metabolic pathways were discovered such as the urea cycle in *T. pseudonana* and the uptake of cyanate as a nitrogen source in *Prochlorococcus* MED4. Up to now, biological oceanographers had not considered cyanate as a nitrogen source of any significance. Genome research may therefore influence paradigms on biogeochemical cycling of elements.

About 50% of sequences from newly sequenced genomes and expressed sequence tags (ESTs) bear no similarity to genes identified previously (Armbrust et al. 2003, Ronning et al. 2003). This proportion may be even larger for organisms from polar environments (Clark et al. 2004). These genes could point to novel physiological and ecological phenomena and may also have the potential for biotechnological applications. The first whole genome sequences from polar microorganisms (e.g. *Colwellia psychrerythraea*, Psychrobacter sp. 273–4) and BAC libraries from ice fishes demonstrate the presence of specific metabolical pathways (http://www.genome.gov/10001852). Recently, the U.S. National Academy of Sciences has mapped a strategy for advancing our understanding of polar regions (Hoag, 2003, National Research Council USA 2003). Genomic sequencing and expression analysis will be at the center of this endeavor.

We selected the pennate diatom *Fragilariopsis cylindrus* for molecular studies on mechanisms of polar adaptation. *Fragilariopsis cylindrus* is regarded as the most important cold water diatom of the polar oceans (von Quillfeldt 2004). This diatom has been reported as a dominant species in sea ice, but can also dominate in open water blooms. The goal of this work was to provide a functionally annotated preliminary set of ESTs from *F. cylindrus* expressed under an important polar environmental condition: freezing temperature. Our results provide the first insights into the genome of a polar diatom and provide evidence for specific sequences that may underlie adaptation to freezing conditions in polar waters.

MATERIALS AND METHODS

Fragilariopsis cylindrus *cDNA library and EST generation*. *Fragilariopsis cylindrus* was isolated from Antarctic sea ice during a "Polarstern" expedition (ANT XVI/3) in the eastern Weddell Sea. Axenic cultures were grown at $+5^\circ$C in a 10 L batch culture under 35 µmol photons $\cdot m^{-2} \cdot s^{-1}$ (16:8 L:D) in double f/2 medium (Guillard and Ryther 1962). Bubbling with air (150 mL/min) ensured sufficient CO_2 supply and continuous mixing. Samples for the cDNA library were taken 5 days after chilling the cells to the freezing point of seawater (approximately -1.8°C). *Fragilariopsis cylindrus* is acclimated to freezing conditions after five days at this irradiance, as shown in an expression study (Mock and Valentin 2004).

Total RNA was isolated with an RNeasy Plant Mini Kit (Qiagen, Hilden, Germany). The mRNA was isolated from approximately 100 µg total RNA with an Oligotex mRNA Midi-Kit (Qiagen). Approximately 800 ng poly A^+ mRNA was used for first-strand cDNA synthesis. The cDNA library was synthesized with a SMART™ cDNA Library Construction Kit (Clontech, Mountain View, CA, USA). Total poly A^+ mRNA was used for first-strand synthesis with SMART IV™ oligonucleotides and CDS III/3′PCR primer. Double-stranded cDNA synthesis was performed by LD PCR with an Eppendorf Thermocycler (Hamburg, Germany) using the following program: 95° C for 5 min denaturation, and subsequent 20 cycles at 95° C (2 min) and 68° C (6 min). The cDNA was digested with *SfiI* and fractionated with CHROMA Spin™-400 columns (Amersham, Braun-Schweigen, Germany). The resulting cDNAs were ligated at 16° C overnight into pTriplEX2 vectors. A separate λ-phage packaging reaction (Promega, Madison, WI, USA) was used to obtain an amplified library with a titer of 2.7×10^9 pfu/mL. Blue/white screening with IPTG and X-gal revealed a recombination efficiency of approximately 70%. We recovered DNA from the clones with the magnetic bead kit from Qiagen. Sequencing of cDNA clones from 5′ end were performed using BigDye terminator chemistry from Applied Biosystems, Foster City, CA, USA. The sequencing reaction products were separated on ABI3700 96 capillary machines. Base calling, vector masking, and sequence quality assessment were performed using Phred (Ewing and Green 1998, Ewing et al. 1998). Sequences with a Phred score less than 20 were rejected from the data set.

EST clustering and assembly analysis. The Phrap algorithm with a minimum quality score of 20 was used for clustering of sequences. Sequence clusters were inspected manually with the help of the Staden package (Staden et al. 1998).

Sequence comparison and functional classification. Individual tentative unique sequences were compared (tBLASTX) to swissprot and genpept data bases on a local Sun system. Furthermore, individual data bases of all available sequences from *Thalassiosira pseudonana* (http://genome.jgi-psf.org/thaps1/thaps1.home.html), *Cyanidioschyzon merolae* (http://merolae.biol.s.u-tokyo.ac.jp/), *Porphyra yezoensis* (http://www.kazusa.or.jp/en/plant/porphyra/EST/), *Oryza sativae* (http://www.tigr.org/tdb/e2k1/osa1/), *Arabidopsis thaliana* (http://www.arabidopsis.org/), *Physcomitrella paten* (http://www.moss.leeds.ac.uk/)s, and *Chlamydomonas reinhardtii* (http://www.chlamy.org/chlamydb.html) were built and queried the same way. Functional domains were searched against COG (http://www.ncbi.nlm.nih.gov/COG/old/xognitor.html) and Interpro (http://www.ebi.ac.uk/interpro/) data bases using default query parameters.

RESULTS

Clustering and assembly analysis. The 5′ ends of 2372 randomly chosen clone inserts were sequenced from a non-normalized cDNA library. Phred analysis (Ew-

TABLE 1. Clustering and redundancy within the cDNA library.

	No. sequences
Total number of ESTs	1376
Singletons	804
ESTs in tentative unique consensi (TCs)	572
TCs	192
Tentative unique sequences	996

EST, expressed sequence tag.

ing et al. 1998) identified 1376 high-quality sequences. To determine the redundancy in the present EST data set and to identify tentative unique sequences, all 1376 ESTs were subjected to sequence clustering and assembly analysis using the Phrap algorithm with standard parameters. Sequence clusters were inspected manually with the help of the Staden package (Staden et al. 1998). The tentative unique sequences consisted of tentative unique singletons and tentative consensus (TC; Ronning et al. 2003) sequences. The ESTs with less than 95% sequence identity over the matching range to other EST sequences from the library were defined as singletons, and TCs were derived from aligned groups of ESTs sharing significant sequence similarity. Approximately 58% of the 1376 *F. cylindrus* ESTs were identified as singletons, whereas approximately 42% of the ESTs were aligned into 192 tentative unique consensus sequences. Finally, a set comprising 996 tentative unique sequences was derived (Table 1). The length of these sequences ranged from 18 to 1580 bp with an average of 415 bp. The number of ESTs in the TCs ranged from two to 31, with an average of 3 ESTs per consensus sequence. The number of ESTs corresponding to the tentative unique sequences (singletons and TCs) ranged from one to 31, with an average of 1.4 ESTs per tentative unique sequence (Fig. 1).

Functional analysis of tentative unique sequences. The 996 tentative unique sequences were compared with 11 non-redundant data bases, of which two (Interpro and COG) were used to identify functional protein domains (supplementary Table 1). Because the *Chlamydomonas* database was composed of a genomic database and an EST database, the total number of databases increased to 12. The comparison was conducted using BLASTX (Altschul et al. 1997) with a cut-off expectancy (e) value of 10^{-4}. Using the distribution of *e*-values for the best matching sequence identified in these 10 data bases, the *F. cylindrus* sequences were most similar to those of the *Thalassiosira* database. Conversely, *Chlamydomonas* sequences showed the lowest similarities to sequences from the *F. cylindrus* data base (Fig. 2). Most of the significant (*e*-value $\leq 10^{-4}$) matches were also retrieved from the *Thalassiosira* data base (total matches: 271). Sixty-nine sequences displayed no similarity to sequences from *Thalassiosira* but were similar to sequences from the

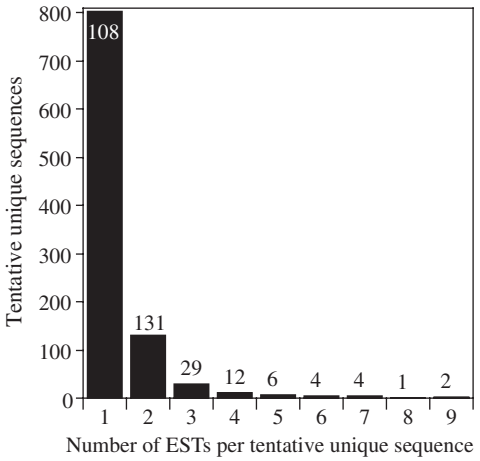

FIG. 1. Distribution of the number of Expressed sequence tags (ESTs) per assembled tentative unique sequence. In total, 1376 EST sequences were analyzed.

other data-bases. A comparison of all matches from all data-bases identified 340 sequences in *F. cylindrus* that were similar to a data-base sequence (Table 2). Interestingly, 84 sequences from *F. cylindrus* had sim-

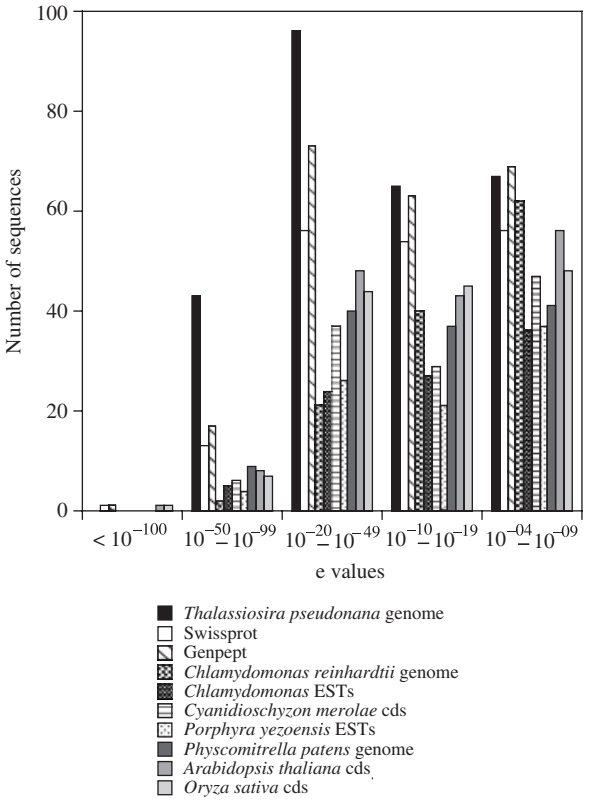

FIG. 2. Distribution of *e*-values from 10 non-redundant databases. Functionally defined database entries were used as resource for assignment of potential functions to the tentative unique sequences. Matches with *e*-values higher than 10^{-10} are most likely insignificant and can may due to e.g. biased nucleotide usage as in *Chlamydomonas reinhardtii*.

TABLE 2. Similarity of the annotated *Fragilariopsis cylindrus* EST database to other databases (cutoff *e*-value $\leq 10^{-4}$).

	996 unique EST sequences from *Fragilariopsis cylindrus*								
Data banks	Total number of similar sequences in data banks	Similar sequences in SwissProt/ GenPep	Similar sequences in *T. pseudonana*	Similar sequences in *C. reinhardtii*	Similar sequences in *C. merolae*	Similar sequences in *P. yezoensis*	Similar sequences in *P. patens*	Similar sequences in higher plants	Similar sequences in *Ectylindrus*
SwissProt/GenPep	229	**17**	179	127	115	77	122	160	229
Thalassiosira pseudonana	271	179	**84**	115	109	70	110	143	271
Chlamydomonas reinhardtii	145	129	115	**3**	93	73	96	115	145
Cyanidioschyzon merolae	119	115	109	93	**0**	64	88	105	119
Porphyra yezoensis	86	77	70	73	64	**2**	67	71	86
Physcomitrella patens	125	122	110	94	88	67	**1**	111	125
Higher plants	173	160	143	115	105	71	111	**3**	173
Fragilariopsis cylindrus	*340*	229	271	147	119	88	127	173	*656*

The table reads as follows: the left and right columns give the total number of similar/homologous sequences found in a data base when compared with the *F. cylindrus* data set, e.g. out of the 996 unique *F. cylindrus* sequences 229 showed similarity to sequences from SwissProt or GenPep, 271 were similar to sequences from *Thalassiosira pseudonana* etc., *340* of the 996 ESTs could be identified by comparison with all databases; *656* could not be identified. The diagonal (in bold) gives the number of unique sequences shared by *F. cylindrus* and a given data base, e.g. 84 of the 996 EST matched only sequences from *T. pseudonana*, only 2 sequences from *Porphyra yezoensis*, etc. The remaining values give the number of sequences shared by three databases, e.g. out of the 996 ESTs, 179 were found in SwissProt/GenPep and in *T. pseudonana*, 93 were found in *C. reinhardtii* plus *C. merolae*.

EST, expressed sequence tag.

FIG. 3. Distribution of matches either in the genome of *Thalassiosira pseudonana* (a) or non-redundant protein data bases (Swissprot and genpept) (b). These sequences from *T. pseudonana*, swissprot, and genpept shared similarity with sequences from either bacteria, fungi, other algae, plants, or animals.

ilar sequences only in *Thalassiosira*, three only in *Chlamydomonas*, two only in *Porphyra*, one only in *Physcomitrella*, three only in higher plants, and 17 only in protein data bases (Swissprot and Genpept) (Table 2, supplementary Table 2). Most of the 84 sequences that occurred exclusively in the genome of *Thalassiosira* showed the highest degree of similarity to eukaryotic counterparts from other algae/plants (30% of sequences), animals (27% of sequences), and fungi (4% of sequences) (Fig. 3a). Twenty-three percent of these sequences were predicted open reading frames (ORFs) with EST support but unknown function (not related to conserved hypothetical proteins), and 16% showed similarity to bacterial sequences. In contrast, matches only found in protein data bases (17 matches) showed the highest degree of similarity to prokaryotic counterparts from bacteria (65% of 17 matches, Fig. 3b). Eighteen percent of these sequences were similar to fungi and 17% to animals. Although bacterial contaminations of the algal culture cannot be ruled out entirely, the relatively high *e*-values (close to 10^{-4}) point to either fast-evolving, horizontally transferred sequences, or to non-significant matches. Sequences found only in *Chlamydomonas*, *Porphyra*, and *Physcomitrella* are related to conserved hypothetical proteins with unknown function, except for one sequence from *Porphyra*, which was identified as an unspecific monooxygenase (*e*-value 4×10^{-23}).

A total of 656 sequences seem to be specific to *F. cylindrus*, a number that may change as more sequences become available from new genome projects (e.g. *Phaeodactylum tricornutum*). In addition, these results from sequence comparison (Table 2 and Fig. 2) are based on heterogenous data sets. For instance, there are many more sequences available from the genome of *Chlamydomonas* than from the genome of *Porphyra*

TABLE 3. Most abundant TCs (tentative unique consensi).

Internal name of TC	Gene definition	No. of clones in sequence
F.cyla04h04.s1	Fucoxanthin chl *a,c*-binding protein	31
F.cyla16se09.s1	Calmodulin-like protein	20
F.cyla19g12.s1	Unknown function; signal peptide predicted	9
F.cyla01c06.s1	Fucoxanthin chl *a,c*-binding protein	9
F.cyla19e03.s1	Unknown function	8
F.cyla19h06.s1	Sm-like protein	7
F.cyla10g01.s1	Unknown function; signal peptide predicted	7
AVIEST.0.231	Unknown function	7
F.cyla08d09.s1	Unknown function; signal peptide predicted	7
F.cyla12e12.s1	Unknown function	6

because the genome of *Chlamydomonas* is about 90% complete, and over 150,000 ESTs (dbEST NCBI) are available. In contrast, genome sequence from *Phorphyra* is not yet available, but approximately 21,000 ESTs (dbEST NCBI) have been sequenced. A comprehensive data bank for *F. cylindrus* ESTs was constructed and can be found at http://genome.imb-jena.de/ALGAE/index.html. Additionally, all ESTs were submitted to http://www.ncbi.nlm.nih.gov/dbEST/.

Most abundant sequences in TCs. Several TCs were composed of mulitple EST sequences (Table 3). The most abundant EST detected encoded a fucoxanthin/chl *a,c* binding protein (fcp) (31 ESTs), followed by a calmodulin-like protein (20 ESTs) and a protein of unknown function with a predicted signal peptide sequence (Interpro data base match). Another abundant fcp (9 ESTs) as well as an Sm-like protein (7 ESTs) could be identified among the 10 most abundant TCs. The remaining five most abundant TCs encoded proteins with unknown function. Two of these TCs had a putative signal peptide sequence followed by an open reading frame (Table 3).

Most abundant protein domains and assigned functional categories. All tentative unique sequences were compared with the Interpro (EBI) and COG (clusters of orthologous groups of proteins, NCBI) data base to identify protein domains. From all tentative unique sequences (996 sequences), 245 sequences (approximately 25%) were found to contain conserved protein domains (supplementary Table 1). The COG database and its functional categories were used to cluster all domains (Table 4). Biological processes such as translation, ribosomal structure, and biogenesis (category J) were represented by 26 protein domains, followed by amino acid transport and metabolism (category E) (17 protein domains), post-translational modifications (category O) (15 protein domains), energy production and conversion (category C) (11 protein domains), and carbohydrate metabolism (category G) (10 protein domains) and general metabolism (category R) (10 protein domains) (Table 4). The most abundant protein domains from both data bases (Interpro, COG) are shown in Table 5.

DISCUSSION

Abundance and function of tentative unique sequences. Only 0.2% (two sequences) of all sequences were highly expressed, i.e. they are represented by TCs comprising more than 10 ESTs (Zhang et al. 2004). A total of 31 ESTs encode a fucoxanthin, chl *a,c*-binding protein (fcp), the major protein components of the light-harvesting antenna complexes of PSI and PSII within diatom plastids (Grossman et al. 1990). In addition, a second gene encoding an fcp was supported with nine ESTs. Some fucoxanthin, chl *a,c* binding proteins are known to be highly expressed in *F. cylindrus*, as indicated by an expression study using macroarrays (Mock and Valentin 2004). A gene encoding an fcp was also supported by the most ESTs in a *P. tricornutum* EST library (Scala et al. 2002), suggesting that members of this gene family (fcps) may be highly expressed under certain stresses or changing environmental conditions. The second TC in this study is composed of 20 ESTs and encodes a calmodulin-like protein whose specific function cannot be

TABLE 4. Most abundant functional categories identified by the COG (cluster of orthologous groups of proteins) data base (http://www.ncbi.nlm.nih.gov/COG/old/xognitor.html).

Rank	Functional category (symbol of category)	No. of protein domains
1	Translation, ribosomal structure, and biogenesis (J)	26
2	Amino acid transport and metabolism (E)	17
3	Post-translational modification, protein turnover, chaperones (O)	15
4	Energy production and conversion (C)	11
5	Carbohydrate transport and metabolism (G)	10
5	General function prediction only (R)	10

TABLE 5. Most abundant protein domains identified by the COG (Cluster of Orthologous Groups of Proteins) (http://www.ncbi.nlm.nih.gov/COG/old/xognitor.html) and Interpro (http://www.edi.ac.uk/interpro/) data bases.

Rank	Description of protein domain	No. of unique sequences
1	Ribosomal	19
2	ABC transporter	10
3	Peptidases	9
4	Chl *a/b* binding	8
5	DNA/RNA helicases	6
5	GTP binding	6
6	Chaperons	5
6	TonB box, N-terminal	5
7	Elongation factor	4
7	Zn-finger	4

defined by sequence comparison. However, of the 996 tentative unique sequences, two IQ calmodulin binding regions (IPR00048) could be identified: one interpro domain matched the annotated calmodulin-like protein and the other a conserved protein with unknown function. Calmodulin acts as a major calcium sensor and orchestrator of regulatory events through its interaction with a diverse group of cellular proteins (Rhoads and Friedberg 1997). Three classes of recognition motifs exist for many of the known calmodulin binding proteins. The IQ motif is a consensus for Ca^{2+}-independent binding, and two related motifs termed 18–14 and 1–5–10 based on the position of conserved hydrophobic residues indicate Ca^{2+}-dependent binding. The calmodulin binding IQ motif occurs in a variety of proteins such as protein kinases, GTPase-activating enzymes, sodium channel proteins and multidrug resistance proteins (Rhoads and Friedberg 1997). The question remains regarding which specific function this protein carries out in *F. cylindrus*.

A moderately expressed sequence (six to nine ESTs per TC) encoded an Sm-like protein. Proteins from the Sm family are known to interact with small mRNAs for mRNA processing (e.g. splicing). The Sm proteins are essential for pre-mRNA splicing and are implicated in the formation of stable, biologically active spliceosomal small nuclear ribonucleoproteins (snRNP) structures that are involved in Sm protein–protein interactions (Hermann et al. 1995).

The remaining six TCs out of the 10 most abundant TCs could not be functionally defined based on analysis of all 12 data bases (tBLASTX and BLASTN), which is unusual based on EST libraries from other eukarytoes (chl Scala et al. 2002, Ronning et al. 2003, Shrager et al. 2003, Habermann et al. 2004, Ida et al. 2004) in which the most highly expressed genes had defined functions. The TCs from *F. cylindrus* were not related to conserved genes with unknown function, nor to genes from other diatoms (*T. pseudonana* and *P. tricornutum*). However, all six TCs had an open reading frame either at the 5′ end or in the middle of the sequence, and five had a polyA tail and a length between 671 and 1263 bp. For 129 ESTs with a polyA tail, we could localize the gene end relative to the polyA tail and were therefore able to calculate the average length of the 3′ UTR, which was 138 bp (5–500 bp). The smallest TC of the six most abundant TCs with unknown function is longer than the longest 3′ UTR sequence. Consequently, it is unlikely that the six abundant unknown TCs are UTRs. They could therefore be specific sequences at least moderately expressed under freezing conditions.

Sequence comparison with data bases. Sixty-six percent of sequences in the EST library from *F. cylindrus* are related to genes of unknown function. However, it is possible that some of these sequences that were shorter (less than 100 bp) and without a polyA tail may have corresponded to UTRs. The remaining 34% of the ESTs showed similarities, expressed as low *e*-values (*e*-value $\leq 10^{-04}$), to sequences from *T. pseudonana* and ESTs from *P. tricornutum*, which were integrated into the annotation of the genome from *T. pseudonana*. This high similarity of sequences between *F. cylindrus* and *T. pseudonana* is related to the fact that they both belong to the class Bacillariophyceae.

Eighty-four EST sequences were similar to sequences found only in the *T. pseudonana* data base but not in the other plant/algae databases examined here. Thirty-one percent of these 84 genes had similarities to genes from heterotrophic eukaryotic organisms (fungi and animals), which were possibly derived from the heterotrophic secondary host, although gene loss in the plant/red algal lineage cannot be ruled out (Armbrust et al. 2003). Twenty-three percent were novel diatom genes with unknown function. Interestingly, all similarities to sequences exclusively found in protein databases (genpept and swissprot) were related to bacteria and other heterotrophs, and none of them were found in the plant databases or the *T. pseudonana* data base examined here. For instance, there are 3 "animal-like" *F. cylindrus* sequences that are absent from *T. pseudonana*: one is most similar to a human calpain protease (PalBH/CAPN7, Acc. Nr. Q9Y6W3), and the other two to the *Dictyostelium discoideum* kinase (SNF1/AMP, Acc. Nr. AAD30963) and MkpA protein (Acc. Nr. AAO51390). Interestingly, all three sequences were related to protein metabolism.

Abundant protein domains and functional distribution. The three most common functional categories, comprised of 58 protein domains (23% of all identified domains), were related to translation, including post-translational modifications and transport of amino acids/peptides. Six different DNA/RNA helicases in the EST library revealed that DNA and RNA coiling/uncoiling are important for this organism. Minimizing secondary structures and duplexes of mRNAs, which could easily form under low temperatures, is necessary to initiate translation. Up-regulation of a DEAD/DEAH box RNA helicase under freezing temperatures already demonstrated the importance of this process in *F. cylindrus* (Mock and

Valentin 2004). However, protein domains of DNA/RNA helicases are the 8th most abundant protein domain in the genome of *T. pseudonana* (Armbrust et al. 2003), and therefore more evidence is necessary to conclude that these enzymes are essential to cope with freezing temperatures.

Nine genes encoding peptidases were identified in the EST library of *F. cylindrus*: two sequences with signal peptidase domains were identified and two with domains for cystein peptidases. Peptidases are enzymes responsible for either the complete digestion of proteins or cleavage of localization peptides required for protein targeting and activation. Three of the 10 most highly expressed sequences with unknown function in *F. cylindrus* possess signal peptides (predicted by Interpro). In addition, peptidases may also be required to repair photodamaged proteins (e.g. D1 of PSII) under freezing temperatures (Mock and Valentin 2004).

Membrane transport of substances other than proteins also seems to play a pivotal role in cold adaptation. The high number of ABC transporters in plants (Henikoff et al. 1997, Sánchez-Fernández et al. 2001) has been hypothesized to reflect the fact that sessile organisms rely upon detoxification as an important means of resisting different stresses. Interestingly, few plant ABC transporters have actually been shown to play a role in detoxification. Instead, these transporters appear to be required for a variety of other processes such as fungal resistance, stomatal conductance, or signal transduction (Martinoia et al. 2002).

Two out of the ten ABC transporter EST sequences, both represented by a single clone, displayed homologies to bacterial permeases. Four of the ABC transporters are homologous to ABC transporters exclusively found in eukaryotes. There are two sequences with homologies to WBC proteins: one PGP-like EST and the other one with similarities to ABC1 proteins. All these proteins are found in animals and plants, and they consist of at least one membrane-spanning domain coupled to an ATP-binding cassette in one polypeptide, like the WBC proteins, or they harbor both domains in tandem. Functional characterizations of these transporters according to their structure were not possible, because of their diverse substrate specificities and functions. Four ABC transporter ESTs isolated in this study are of particular interest. They are all homologous to genes encoding proteins with two ATP-binding cassettes without membrane spanning domains. One of them belongs to the GCN20 class, first characterized in yeast and then also found in other higher eukaryotes (Vazquez de Aldana et al. 1995, Dean and Allikmets 2001, Sánchez-Fernández et al. 2003). Another two ESTs encode two different YEF3 homologs. These proteins are structurally and functionally related to GCN20 proteins in yeast. Both protein classes are involved in translational control in yeast (Decottignies and Goffeau 1997), but their function in other organisms remains to be elucidated. The fourth EST clone shows homology to bacterial *uup* genes and also to a protein with two ABC domains but without a membrane-spanning domain. These proteins have been reported to control transpositional processes in *E. coli* (Reddy and Gowrishankar 1997). Although all these structurally related proteins are relatively well characterized in bacteria and yeast, almost nothing is known about them in higher eukaryotes. The appearance of these four different clones in this EST collection suggests a functional role for this particular group of ABC transporters in cold acclimation of *F. cylindrus*.

To date, polar diatoms are not a major subject in the field of cold acclimation/adaptation, despite their important role as the basis of the entire polar food web. Therefore, this EST study was conducted to provide the basis for further molecular studies with polar diatoms and in particular with *F. cylindrus*. Most of the annotated sequences are related to translation, ribosomal structure, biogenesis, and post-translational modifications of proteins. New enzymes/proteins are required to acclimate to freezing temperatures. The occurrence of signal peptides in some of the most highly expressed sequences indicates secretion of these proteins.

We are grateful for the constructive advice from two anonymous referees, Chris Bowler and particularly, Ginger Armbrust, who helped us with the final editing of this paper.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–402.

Armbrust, E. V., Berges, J. A., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., Zhou, S., Allen, A. E., Apt, K. E., Bechner, M., Brzezinski, M. A., Chaal, B. K., Chiovitti, A., Davis, A. K., Demarest, M. S., Detter, J. C., Glavina, T., Goodstein, D., Hadi, M. Z., Hellsten, U., Hildebrand, M., Jenkins, B. D., Jurka, J., Kapitonov, V. V., Kröger, N., Lau, W. W., Lane, T. W., Larimer, F. W., Lippmeier, J. C., Lucas, S., Medina, M., Montsant, A., Obornik, M., Parker, M. S., Palenik, B., Pazour, G. J., Richardson, P. M., Rynearson, T. A., Saito, M. A., Schwartz, D. C., Thamatrakoln, K., Valentin, K., Vardi, A., Wilkerson, F. P. & Rokhsar, D. S. 2003. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* 306:79–86.

Boyd, P. W. 2002. Environmental factors controlling phytoplankton processes in the Southern Ocean. *J. Phycol.* 38:844–61.

Clark, M. S., Clarke, A., Cockell, C. S., Convey, P., Detrich III, H. W., Fraser, K. P. P., Johnston, I. A., Methe, B. A., Murray, A. E., Peck, L.S, Römisch, K. & Rogers, A. D. 2004. Antarctic genomics. *Comp. Funct. Genom.* 5:230–8.

Coale, K. H., Johnson, K. S., Chavez, F. P., Buesseler, K. O., Barber, R. T., Brzezinski, M. A., Cochlan, W. P., Millero, F. M., Falkowski, P. G., Bauer, J. E., Wanninkhof, R. H., Kudela, R. M., Altabet, M. A., Hales, B. E., Takahashi, T., Landry, M. R., Bidigare, R. R., Wang, X., Chase, Z., Strutton, P. G., Friederich, G. E., Gorbunov, M. Y., Lance, V. P., Hilting, A. K., Hiscock, M. R., Demarest, M., Hiscock, W. T., Sullivan, K. F., Tanner, S. J., Gordon, R. M., Hunter, C. N., Elrod, V. A., Fitzwater, S. E., Jones, J. L., Tozzi, S., Koblizek, M., Roberts, A. E., Herndon, J., Brewster, J., Ladizinsky, N., Smith, G., Cooper, D., Timothy, D., Brown, S. L., Selph, K. E., Sheridan, C. C., Twining, B. S. & Johnson, Z. I. 2004. Southern ocean iron enrichment experiment: carbon cycling in high- and low-Si waters. *Science* 304:408–14.

Cota, G. F. 1985. Photoadaptation of high Arctic ice algae. *Nature* 315:219–22.

Dean, M. & Allikmets, R. 2001. Complete characterization of the human ABC gene family. *J. Bioenerg. Biomembr.* 33: 475–9.

Decottignies, A. & Goffeau, A. 1997. Complete inventory of the yeast ABC proteins. *Nat. Genet.* 15:137–45.

Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I. M., Barbe, V., Duprat, S., Galperin, M. Y., Koonin, E. V., Gall, F., Makarova, K. S., Ostrowski, M., Oztas, S., Robert, C., Rogozin, I. B., Scanlan, D. J., Tandeau, N., Weissenbach, J., Wincker, P., Wolf, Y. I. & Hess, W. R. 2003. Genome sequence of the cyanobacterium Prochlorococcus marinus SS120, a nearly minimal oxyphototrophic genome. *Proc. Nat. Acad. Sci.* 100:10020–5.

Ewing, B. & Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8:186–94.

Ewing, B., Hillier, L., Wendel, M. C. & Green, P. 1998. Base-calling of automated sequencer traces using phred: I. Accuracy assessment. *Genome Res.* 8:175–85.

Falkowski, P. G. & Davis, C. S. 2004. Natural proportions. *Nature* 431:131.

Falkowski, P. G., Katz, M. E., Knoll, A. H., Quigg, A., Raven, J.A, Schofield, O. & Taylor, F. J. 2004. The evolution of modern eukaryotic phytoplankton. *Science* 305:354–60.

Fiala, M. & Oriol, L. 1990. Light-temperature interactions on the growth of Antarctic diatoms. *Polar Biol.* 10:629–36.

Grossman, A., Manodori, A. & Snyder, D. 1990. Light-harvesting proteins of diatoms: their relationship to the chlorophyll a/b binding proteins of higher plants and their mode of transport into plastids. *Mol. Gen. Genet* 224:91–100.

Guillard, R. R. & Ryther, J. H. 1962. Studies of marine plankton diatoms. I. *Cyclotella nana* (Husted) and *Detonula confervacea* (Cleve). *Can. J. Microbiol.* 8:229–39.

Habermann, B., Bebin, A. G., Herklotz, S., Volkmer, M., Eckelt, K., Pehlke, K., Epperlein, H. H., Schackert, H. K., Wiebe, G. & Tanaka, E. M. 2004. An *Ambystoma mexicana* EST sequencing project: analysis of 17,352 expressed sequence tags from embryonic and regenerating blastema cDNA libraries. *Genome Biol.* 5:R67.

Henikoff, S., Greene, E. A., Pietrokovski, S., Bork, P., Attwood, T. E. & Hood, L. 1997. Gene families: the taxonomy of protein paralogs and chimeras. *Science* 278:609–14.

Hermann, H., Fabrizio, P., Raker, V. A., Foulaki, K., Hornig, H., Brahms, H. & Luhrmann, R. 1995. snRNP Sm proteins share two evolutionarily conserved sequence motifs which are protein-protein interactions. *EMBO J.* 14:2076–88.

Hoag, H. 2003. Genomes take pole position in icy wastes. *Nature* 421:880.

Ida, H., Boylan, S. A., Weigel, A. L., Smit-McBride, Z., Chao, A., Gao, J., Buchoff, P., Wistow, G. & Hjelmeland, L. M. 2004. EST analysis of mouse retina and RPE/choroid cDNA libraries. *Mol. Vis.* 10:439–44.

Legendre, L., Ackley, S. F., Dieckmann, G. S., Gulliksen, B., Horner, R., Hoshiai, T., Melnikov, I. A., Reeburgh, W. S., Spindler, M. & Sullivan, C. W. 1992. Ecology of sea ice biota 2. Global significance. *Polar Biol.* 12:429–44.

Leventer, A. 1998. The fate of Antarctic "sea ice diatoms" and their use as paleoenvironmental indicators. *In* Lizotte, M. P. & Arrigo, K. R. [Eds.] *Antarctic Sea Ice, Biological Processes, Interactions and Variability. Antarctic Research Series*. Vol. 73. American Geophysical Union, Washington, DC, USA, pp. 121–37.

Lizotte, M. P. 2001. The contribution on sea ice algae to Antarctic marine primary production. *Amer. Zool.* 41:57–73.

Martinoia, E., Klein, M., Geisler, M., Bovet, L., Forestier, C., Kolukisaoglu, U., Müller-Röber, B. & Schulz, B. 2002. Multifunctionality of plant ABC transporters–more than just detoxifiers. *Planta* 214:345–55.

Mock, T. & Valentin, K. 2004. Photosynthesis and cold acclimation: molecular evidence from a polar diatom. *J. Phycol.* 40: 732–41.

NRC. 2003. *Frontiers in Polar Biology in the Genomic Era*. National Academy Press, Washington, DC (http://www.nap.edu/catalog/10623.html).

Palenik, B., Brahamsha, B., Larimer, F. W., Land, M., Hauser, L., Chain, P., Lamerdin, J., Regala, W., Allen, E. E., McCarren, J., Paulsen, I., Dufresne, A., Partensky, F., Webb, E. A. & Waterbury, J. 2003. The genome of a motile marine Synechococcus. *Nature* 424:1037–42.

Quillfeldt, C. H. 2004. The diatom *Fragilariopsis cylindrus* and its potential as an indicator species for cold water rather than for sea ice. *Vie Milieu* 54(2–3):137–43.

Reddy, M. & Gowrishankar, J. 1997. Identification and characterization of *ssb* and *uup* mutants with increased frequency of precise excision of transposon Tn*10* derivatives: Nucleotide sequence of *uup* in *Escherichia coli*. *J. Bacteriol.* 179:2892–9.

Rhoads, A. R. & Friedberg, F. 1997. Sequence motifs for calmodulin recognition. *FASEB J.* 11:331–40.

Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A., Arellano, A., Coleman, M., Hauser, L., Hess, W. R., Johnson, Z. I., Land, M., Lindell, D., Post, A. F., Regala, W., Shah, M., Shaw, S. L., Steglich, C., Sullivan, M. B., Ting, C. S., Tolonen, A., Webb, E. A., Zinser, E. R. & Chisholm, S. W. 2003. Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature* 424:1042–7.

Ronning, C. M., Stegalkina, S. S., Ascenzi, R. A., Bougri, O., Hart, A. L., Utterbach, T. R., Vanaken, S. E., Riedmuller, S. B., White, J. A., Cho, J., Pertea, G. M., Lee, Y., Karamycheva, S., Sultana, R., Tsai, J., Quackenbush, J., Griffiths, H. M., Restrepo, S., Smart, C. D., Fry, W. E., Hoeven, R., Tanksley, S., Zhang, P., Jin, H., Yamamoto, M. L., Baker, B. J. & Buell, C. R. 2003. Comparative analysis of potato expressed sequence tag libraries. *Plant Physiol.* 131:419–29.

Sánchez-Fernández, R., Davies, T. G. E., Coleman, J. O. D. & Rea, P. A. 2001. The Arabidopsis thaliana ABC protein superfamily, a complete inventory. *J. Biol. Chem.* 276:30231–44.

Scala, S., Carels, N., Falciatore, A., Chiusano, M. L. & Bowler, C. 2002. Genome proterties of the diatom *Phaeodactylum tricornutum*. *Plant Physiol.* 129:993–1002.

Shrager, J., Hauser, C., Chang, C. W., Harris, E. H., Davies, J., McDermott, J., Tamse, R., Zhang, Z. & Grossman, A. R. 2003. *Chlamydomonas reinhardtii* genome project. A guide to the generation and use of the cDNA information. *Plant Physiol.* 131:401–8.

Smetacek, V. 1999. Diatoms and the ocean carbon cycle. *Protist* 150:25–32.

Staden, R., Beal, K. F. & Bonfield, J. K. 1998. The Staden Package, Computer Methods in Molecular Biology. *In* Misener, S. & Krawetz, S. A. [Eds.] *Bioinformatics Methods and Protocols*. vol. 132. The Humana Press Inc., Totowa, NJ, pp. 115–30.

Thomas, D. N. & Dieckmann, G. S. 2002. Antarctic sea ice—a habitat for extremophiles. *Science* 295:641–4.

Vazquez de Aldana, C. R., Marton, M. J. & Hinnebusch, A. G. 1995. GCN20, a novel ATP binding cassette protein, and GCN1 reside in a complex that mediates activation of the eIF-2 alpha kinase GCN2 in amino acid-starved cells. *EMBO J.* 14: 3184–99.

Zhang, H., Sreenivasulu, N., Weschke, W., Stein, N., Rudd, S., Radchuk, V., Potokina, E., Scholz, U., Schweizer, P., Zierold, U., Langridge, P., Varshney, R. K., Wobus, U. & Graner, A. 2004. Large-scale analysis of the barley transcriptome based on expressed sequence tags. *Plant J.* 40:276–90.

Supplementary Material

The following supplementary material is available for this article online:

Table S1. COG and IPR IDs.
Table S2. Similarity of ESTs to other data bases.